

BACKWARD STEP CONTROL FOR HILBERT SPACE PROBLEMS*

ANDREAS POTSCHKA†

Abstract. We extend backward step control for the globalization of convergence of Newton-type methods for nonlinear root finding problems from the finite dimensional to a Hilbert space setting. We show that the method itself can be transferred as is and that subtle adaptations in the required assumptions lead to an equally powerful convergence theory as in the finite dimensional case. The results include global convergence to a distinctive solution characterized by propagating the initial guess by a generalized Newton flow with guaranteed bounds on the discrete nonlinear residual norm decrease and an (also numerically) easily controllable asymptotic linear residual convergence rate. The convergence theory can be exploited to construct efficient numerical methods, which we demonstrate for the case of a Krylov–Newton method and an approximation-by-discretization framework. Both approaches optimize the asymptotic linear residual convergence rate, either over the Krylov subspace or through adaptive discretization, which in turn yields practical and efficient stopping criteria and refinement strategies that balance the linearization errors with the approximation errors of the linear systems. We apply these methods to the class of nonlinear elliptic boundary value problems and present numerical results for the Carrier equation and the minimum surface equation.

Key words. Newton-type methods, globalization, Hilbert space, backward step control

AMS subject classifications. 65J15, 58C15, 65F08, 35J66, 74S05

1. Introduction. The goal of this paper is to extend the theory and numerics of backward step control [24] for nonlinear root finding problems from the finite dimensional setting to an infinite dimensional Hilbert space setting. As a byproduct, we obtain a more elegant proof of convergence also for the special case of finite dimensional problems.

To this end, let U be a Banach space with norm $\|\cdot\|_U$ and V be a Hilbert space (we discuss generalizations to Banach spaces in §2.8) with inner product $(\cdot, \cdot)_V$ and norm $\|v\|_V = \sqrt{(v, v)_V}$ for $v \in V$. For some open subset $D \subseteq U$, let $F : D \rightarrow V$ be continuously Fréchet-differentiable with derivative $F' : D \rightarrow \mathcal{L}(U, V)$, where $\mathcal{L}(U, V)$ denotes the Banach space of all bounded linear operators from U to V . We consider the problem of finding an unknown $u \in D$ such that

$$(1) \quad F(u) = 0_V$$

with a Newton-type iteration: Given $u_0 \in D$, find a suitable approximation of the inverse $M : U \rightarrow \mathcal{L}(V, U)$ of $F'(u)$ and a step size sequence $(t_k)_{k \in \mathbb{N}}$ satisfying $t_k \in [0, 1]$ such that the iteration

$$(2) \quad u_{k+1} = u_k + t_k \delta u_k \quad \text{with } \delta u_k = -M(u_k)F(u_k)$$

converges to a solution $u^* \in D$ of (1). The first and main part of this article is devoted to finding a suitable step size sequence $(t_k)_{k \in \mathbb{N}}$ in §2. Out of the many ways to construct $M(u)$, we elaborate on two choices in §3. For convenience, we define $f : D \rightarrow U$ as

$$f(u) = M(u)F(u).$$

*Submitted to the editors August 5, 2016.

Funding: This work was funded by the European Research Council through S. Engell’s and H.G. Bock’s ERC Advanced Investigator Grant MOBOCON (291 458) and by the German Federal Ministry of Education and Research under grant 05M2013.

†Interdisciplinary Center for Scientific Computing, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany (potschka@iwr.uni-heidelberg.de).

As in [24], our convergence analysis will be based on generalized Newton paths $u^k : [0, \infty) \rightarrow U$, which are defined as the solutions of the initial value problems

$$(3) \quad \frac{du^k}{dt}(t) = -f(u^k(t)) \quad \text{for } t \in [0, \infty) \quad \text{with } u^k(0) = u_k.$$

We shall prove existence and uniqueness of solutions to (3) in our setting in Theorem 12. Note that (2) is an Explicit Euler discretization of (3) with step sizes $(t_k)_{k \in \mathbb{N}}$.

In the theory and the numerics below, the operator M does not appear explicitly anymore and only the function f will be required, which implicitly defines $M(u)$ in the direction $F(u)$. As it turns out, all other directions of $M(u)$ are not important.

The convergence theory below lends itself immediately to the construction of numerical algorithms for the solution to (1) via (2). In particular, it can be used to construct $M(u)$ from $F'(u)$ by finite dimensional approximation, which can then be exploited to construct adaptive discretization schemes that optimize the contraction rate of the algorithm in V . In the case of Finite Element analysis, our approach delivers an adaptive mesh refinement algorithm that can be used straight-forward as another tool complementing refinement strategies based on a posteriori error estimation (see, e.g., [16, 1, 6]). Our approach allows the coupling of discretization and linearization error and the derivation of balanced stopping criteria for the linear problems as in [25], while requiring no formulation of dual problems.

Contributions. In this article, we extend the convergence analysis of backward step control for (2) from the finite dimensional to the Hilbert space setting. We provide reasonable assumptions and provide convergence results for (2) with backward step control. The main result is convergence to a distinct solution characterized by the propagation of the initial guess by the generalized Newton flow (3) provided that no singularity of the problem interferes. In addition, we prove an a priori bound on the nonlinear reduction of the residual norm. The convergence theory can be exploited to construct efficient numerical algorithms, which we discuss for the case of a GMRES–Newton method and a Finite Element approximation. Both are based on the optimization of the residual contraction constant, which yields in the latter case an efficient adaptive mesh refinement strategy. The results are demonstrated for the numerical solution of the Carrier equation and the minimal surface equation.

Overview. In §2, we discuss the general assumptions, explain the method of backward step control, provide reasons why alternative time stepping methods for (3) are not advisable, provide step size bounds, and establish the notion of generalized Newton paths, which are the central ingredient for the analysis of local and global convergence of backward step control. We then discuss extensions to Banach spaces and present an algorithmic realization. In §3, we exploit the convergence analysis to construct two Newton-type methods, a Krylov–Newton method and a method based on approximation-by-discretization. We apply these methods in §4 to the class of nonlinear elliptic boundary value problems and provide numerical results for the Carrier equation and the minimum surface equation in §5.

Notation. We denote the open ball of radius $r > 0$ around $u \in U$ by $B(u, r)$ and the Laplace operator by $\Delta = \nabla \cdot \nabla$. As usual, we write C^0 for the space of continuous functions, $H_0^1(\Omega)$ for the Sobolev space of square integrable functions on a bounded domain $\Omega \subset \mathbb{R}^n$ that vanish at the boundary and admit square integrable derivatives, and $H^{-1}(\Omega)$ for its dual space. The Euler number is denoted by $e = \sum_{k=0}^{\infty} \frac{1}{k!}$.

Acknowledgments. The author is grateful to Felix Lenders for his comments on an earlier draft of this manuscript.

2. Convergence analysis. The overall structure of the backward step control convergence analysis in Hilbert spaces is similar to the finite dimensional case [24]. The intricate interplay of the changes in the details, however, advises us to present the convergence analysis in a self-contained fashion.

2.1. Discussion of assumptions. We start with the following definitions.

DEFINITION 1. The level function $T : D \rightarrow \mathbb{R}$ is $T(u) = \frac{1}{2} \|F(u)\|_V^2$.

DEFINITION 2. The level set of $u \in D$ is $\tilde{\mathcal{T}}(u) = \{\bar{u} \in D \mid T(\bar{u}) \leq T(u)\}$.

DEFINITION 3. The path connected level set of $u \in D$ is

$$\mathcal{T}(u) = \left\{ \bar{u} \in \tilde{\mathcal{T}}(u) \mid \exists c \in C^0([0, 1], \tilde{\mathcal{T}}(u)) \text{ with } c(0) = u, c(1) = \bar{u} \right\}.$$

DEFINITION 4. For $r \in (1, \infty)$ the set of r -regular points is

$$\mathcal{R}_r = \{u \in D \mid r^{-1} \|F(u)\|_V < \|f(u)\|_U < r \|F(u)\|_V\}.$$

DEFINITION 5. The set of ∞ -regular points is $\mathcal{R}_\infty = \bigcup_{r \in (1, \infty)} \mathcal{R}_r$.

We remark that if $u \in D \setminus \mathcal{R}_\infty$, which means that $u \notin \mathcal{R}_r$ for all $r \in (1, \infty)$, then $M(u)$ is either not bounded or does not admit a bounded inverse [28, §I.6, Corollaries 2, 3]. The contrary is, however, not true: $M(u)$ may be unbounded or not admit a bounded inverse although $u \in \mathcal{R}_r$ for some $r \in (1, \infty)$, because in the definition of \mathcal{R}_r only the action of $M(u)$ in direction $F(u)$ is of interest.

We require the following assumptions to hold true:

A1. There exists an $r \in (1, \infty)$ such that $u_0 \in \mathcal{R}_r$, and $\|F(u_0)\|_V > 0$.

A2. There exists a $\kappa < 1$ such that

$$\|F(u) - F'(u)f(u)\|_V \leq \kappa \|F(u)\|_V \quad \text{for all } u \in \mathcal{R}_r \cap \mathcal{T}(u_0).$$

A3. There exists an $\omega < \infty$ such that

$$\|[F'(u) - F'(u - tf(u))]f(u)\|_V \leq \omega t \|f(u)\|_U \|F(u)\|_V \quad \text{for all } u \in \mathcal{T}(u_0), t \in [0, 1].$$

A4. There exists an $L < \infty$ such that

$$\|f(u) - f(\bar{u})\|_U \leq L \|u - \bar{u}\|_U \quad \text{for all } u, \bar{u} \in \mathcal{T}(u_0).$$

A5. For all $\eta > 0$ there exist constants $\gamma, t_\gamma > 0$ such that

$$\|f(u - tf(u)) - f(u)\|_U \geq \gamma t \quad \text{for all } t \in [0, t_\gamma], u \in \mathcal{R}_r \cap \mathcal{T}(u_0) \text{ with } \|f(u)\|_U > \eta.$$

The main difference in the assumptions here compared to the finite dimensional setting in [24] is the weakening of A2 and A3 from a formulation with matrices to a formulation which requires the properties to hold only in the direction of the residual $F(u)$. Thus, all requirements can be postulated without using norms for operators that map between U and V . Apart from the avoidance of operator norms, we had to replace all arguments based on compactness of bounded sets by other means for the proofs in the Hilbert space case. The discussion of the assumptions in [24] still applies to a large extent here: We require in A1 that u_0 is an r -regular point but not a solution. The central κ -condition A2 is a contravariant version of Bock's covariant κ -condition [9] (for a discussion of different affine invariances see [10]). The ω -condition A3 measures a combination of the nonlinearity and the well-posedness of (1) because if F' is Lipschitz continuous with Lipschitz constant L' , then we obtain

$$\|[F'(u) - F'(u - tf(u))]f(u)\|_V \leq L' t \|f(u)\|_U^2$$

and boundedness of $M(u)$ in direction $F(u)$ with constant C implies A3 with $\omega = CL'$. The Lipschitz condition A4 is classical. The nonstandard assumption A5 follows, for instance, if f is bi-Lipschitz with constant ℓ

$$\|f(u - tf(u)) - f(u)\|_U \geq \ell t \|f(u)\|_U$$

with $\gamma = \eta\ell$ and t_γ arbitrary.

2.2. Backward step control. In order to determine t_k , we consider the backward iterate

$$\bar{u}_k(t_k) := u_{k+1} + t_k f(u_{k+1}) = u_k + t_k g(u_k, t_k) \quad \text{with } g(u, t) := f(u - tf(u)) - f(u).$$

The point $\bar{u}_k(t_k)$ is the starting point of an implicit Euler step for (3) that arrives exactly at u_{k+1} , the result of an explicit Euler step starting from u_k . The idea of backward step control is to bound the distance between u_k and $\bar{u}_k(t_k)$ by requiring for some fixed $H > 0$

$$(\text{BSC}) \quad t_k = \min \mathcal{B}_H(u_k) \quad \text{where } \mathcal{B}_H(u) = \{t \in [0, 1] \mid H = t \|g(u, t)\|_U\} \cup \{1\},$$

which implies $\|\bar{u}_k(t_k) - u_k\|_U \leq H$ (with equality for $t_k < 1$) by continuity of g .

2.3. Alternative time stepping methods. The question whether explicit Euler is really the best method to solve (3) arises naturally. We can answer this question affirmatively for two reasons: First, all implicit methods have the drawback that an approximated inverse of an operator involving derivatives of the approximated inverse $M(u)$ would be required, e.g., in the case of the implicit Euler method

$$0 = u_{k+1} + t_k f(u_{k+1}) - u_k,$$

with a local Newton corrector

$$u_{k+1}^{i+1} = u_{k+1}^i - [I_U + t_k f'(u_{k+1}^i)]^{-1} (u_{k+1}^i + t_k f(u_{k+1}^i) - u_k),$$

which is not readily available and would require higher regularity of M than guaranteed by the assumptions above. Second, higher order methods would destroy the well-known locally quadratic convergence of the Newton method, where $M(u) = (F'(u))^{-1}$. This can be seen from the homotopy formulation

$$(4) \quad F(u(t)) - e^{-t} F(u_0) = 0,$$

which we can differentiate with respect to t to arrive exactly at (3) provided that $F'(u(t))$ stays invertible. Thus, the second order truncation error of explicit Euler is required to obtain locally quadratic convergence, because higher consistency orders would result in the locally linear convergence dictated by (4).

2.4. Step size bounds.

LEMMA 6. *If A1 and A4 hold, then (BSC) delivers full steps $t_k = 1$ in the vicinity of a solution $u^* \in \mathcal{R}_r \cap \mathcal{T}(u_0)$.*

Proof. Let $u_k \in B(u^*, L^{-2}H)$. Hence, it holds for all $t \in [0, 1]$ that

$$t \|g(u_k, t)\|_U \stackrel{\text{A4}}{\leq} Lt^2 \|f(u_k)\|_U = Lt^2 \|f(u_k) - f(u^*)\|_U \stackrel{\text{A4}}{\leq} L^2 t^2 \|u_k - u^*\|_U < H.$$

Thus, $\mathcal{B}_H(u_k) = \{1\}$ and $t_k = 1$ by virtue of (BSC). \square

LEMMA 7. If A1 and A4 hold, then (BSC) generates for all $u_k \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ step sizes that are either $t_k = 1$ or have the lower bounds

$$t_k \geq \frac{\sqrt{H}}{\sqrt{L} \|f(u_k)\|_U} > \frac{\sqrt{H}}{\sqrt{rL} \|F(u_k)\|_V} \geq \frac{\sqrt{H}}{\sqrt{rL} \|F(u_0)\|_V}.$$

Proof. If $t_k < 1$, then

$$t_k^2 \stackrel{\text{(BSC)}}{=} \frac{t_k H}{\|g(u_k, t_k)\|_U} \stackrel{\text{A4}}{\geq} \frac{H}{L \|f(u_k)\|_U} \stackrel{\text{A1}}{>} \frac{H}{rL \|F(u_k)\|_V} \geq \frac{H}{rL \|F(u_0)\|_V}. \quad \square$$

LEMMA 8. Let A1 and A5 hold and let $\bar{t} \in (0, 1)$ and $\eta > 0$. Then there exists an $\bar{H} > 0$ such that for all $H \in (0, \bar{H}]$ and $u \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ with $\|f(u)\|_U \geq \eta$ it holds that $\min \mathcal{B}_H(u) \leq \bar{t}$.

Proof by contradiction. We assume to the contrary that for all $\bar{H} > 0$ there exists an $H \in (0, \bar{H}]$ and a $u \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ satisfying $\|f(u)\|_U \geq \eta$ and $\min \mathcal{B}_H(u) > \bar{t}$. Then, A5 guarantees the existence of $\gamma, t_\gamma > 0$ such that for $t := \min\{t_\gamma, \bar{t}\} < \min \mathcal{B}_H(u)$ we obtain from (BSC) that

$$\bar{H} \geq H \geq t \|g(u, t)\|_U \geq \gamma t^2 > 0.$$

Because η and thus γ and t are independent of \bar{H} , we obtain a contradiction for $\bar{H} \rightarrow 0$. \square

LEMMA 9. Let A1, A4, and A5 hold and let $\eta > 0$. Then there exists an $\bar{H} > 0$ such that for all $H \in (0, \bar{H}]$ and $u \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ it holds that

$$\|f(u)\|_U \min \mathcal{B}_H(u) \leq \eta.$$

Proof. We choose $\bar{t} \in (0, 1)$ sufficiently small so that it satisfies $r \|F(u_0)\|_V \bar{t} \leq \eta$. Then, Lemma 8 yields the existence of an $\bar{H} > 0$ such that for all $H \in (0, \bar{H}]$ and all $u \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ with $\|f(u)\|_U \geq \eta$ it holds that $\min \mathcal{B}_H(u) \leq \bar{t}$. Hence,

$$\|f(u)\|_U \min \mathcal{B}_H(u) \stackrel{\text{A1}}{<} r \|F(u)\|_V \bar{t} \leq r \|F(u_0)\|_V \bar{t} \leq \eta.$$

For the remaining $u \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ the assertion holds by virtue of $\|f(u)\|_U < \eta$. \square

2.5. Finite arclength of generalized Newton paths. In the next step, we study the generalized Newton paths given by (3).

LEMMA 10. If A1, A2, A4 and $u_k \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ hold, then there exists $\bar{t} > 0$ such that (3) has a unique local solution $u^k(t) \in \mathcal{R}_r \cap \mathcal{T}(u_k)$ for $t \in [0, \bar{t})$ which satisfies

$$\|F(u^k(t))\|_V \leq e^{-(1-\kappa)t} \|F(u_k)\|_V \quad \text{for all } t \in [0, \bar{t}).$$

Proof. The Picard-Lindelöf theorem [2, II.7, exercise 3] yields with A4 the existence of a unique local solution $u^k(t)$ to (3) in some neighborhood $(-\bar{t}, \bar{t})$ of $t = 0$. Without loss of generality, $\bar{t} > 0$ is small enough to ensure $u^k(t) \in \mathcal{R}_r$ for $t \in [0, \bar{t})$ because \mathcal{R}_r is open. For ease of notation, we abbreviate $u^k(t)$ by u . The Cauchy-Schwarz inequality and A2 show that the level function is nonincreasing along this solution because

$$\begin{aligned} \frac{d}{dt} T(u) &= \left(F(u), F'(u) \frac{du}{dt} \right)_V = - (F(u), F'(u) f(u))_V \\ &= - \|F(u)\|_V^2 + (F(u), F(u) - F'(u) f(u))_V \\ &\leq - \|F(u)\|_V^2 + \kappa \|F(u)\|_V^2 = -2(1 - \kappa) T(u) \leq 0. \end{aligned}$$

Gronwall's inequality (see, e.g., [2]) yields

$$T(u^k(t)) \leq e^{-2(1-\kappa)t} T(u_k)$$

and thus $u^k(t) \in \mathcal{T}(u_k)$ for $t \in [0, \bar{t})$. The assertion follows after multiplication by two and taking square roots. \square

We show in Theorem 12 below that the quantities in the following definition are well-defined under suitable assumptions.

DEFINITION 11. *For $r \in (1, \infty)$, we define the r -regular part u_r^k of the generalized Newton path u^k as the solution to the initial value problem*

$$\frac{du_r^k}{dt}(t) = \begin{cases} -f(u_r^k(t)) & \text{for } u_r^k(t) \in \mathcal{R}_r, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } t \in [0, \infty), \quad \text{with } u_r^k(0) = u_k.$$

We denote its limit by $u_k^* = \lim_{t \rightarrow \infty} u_r^k(t)$ and define $t_k^* = \inf\{t \in [0, \infty) \mid u_r^k(t) \notin \mathcal{R}_r\}$ with the usual convention that $\inf \emptyset = \infty$.

THEOREM 12. *Let A1, A2, and A4 hold. If $u_k \in \mathcal{R}_r \cap \mathcal{T}(u_0)$, then the r -regular part of the generalized Newton path exists uniquely and has a finite arclength satisfying*

$$\|u_k - u_k^*\|_U \leq \ell(u_r^k) < \frac{r}{1-\kappa} \|F(u_k)\|_V < \frac{r^2}{1-\kappa} \|f(u_k)\|_U.$$

If $u_k^* \in \mathcal{R}_r$, then $F(u_k^*) = 0$.

Proof. The unique local solution of Lemma 10 can be extended uniquely in $\mathcal{T}(u_k)$ by repeated application of the Picard–Lindelöf theorem either until $u^k(t) \notin \mathcal{R}_r$ for some $t = t_k^*$ or to the whole interval $t \in [0, \infty)$. In the first case, the r -regular part is uniquely determined by $u_r^k(t) = u_r^k(t_k^*)$ for all $t \geq t_k^*$. We can now use the definition of \mathcal{R}_r and Lemma 10 in order to show

$$\begin{aligned} \ell(u_r^k) &= \int_0^\infty \left\| \frac{du_r^k}{dt}(t) \right\|_U dt = \int_0^{t_k^*} \|f(u^k(t))\|_U dt < r \int_0^{t_k^*} \|F(u^k(t))\|_V dt \\ &\leq r \int_0^\infty e^{-(1-\kappa)t} dt \|F(u_k)\|_V = \frac{r}{1-\kappa} \|F(u_k)\|_V < \frac{r^2}{1-\kappa} \|f(u_k)\|_U. \end{aligned}$$

We obtain the lower arclength bound by noting that the shortest path between u_k and u_k^* has arclength $\|u_k - u_k^*\|_U$. If $u_k^* \in \mathcal{R}_r$, then $t_k^* = \infty$ and Lemma 10 reveals

$$\|F(u_k^*)\|_V \leq \lim_{t \rightarrow \infty} e^{-(1-\kappa)t} \|F(u_k)\|_V = 0. \quad \square$$

2.6. Local convergence. As a prerequisite, we prove that every neighborhood of an isolated zero u^* of F contains a path connected level set that contains a neighborhood of u^* .

LEMMA 13. *Let A1, A2, and A4 hold. If there exist $\varepsilon > 0$ and $u^* \in D$ such that u^* is the only zero of F on $B(u^*, \varepsilon) \subseteq \mathcal{R}_r \cap \mathcal{T}(u_0)$, then there exists an $\tilde{\varepsilon} > 0$ with*

$$\bigcup_{u \in B(u^*, \tilde{\varepsilon})} \mathcal{T}(u) \subseteq B(u^*, \varepsilon).$$

Proof by contradiction. We assume to the contrary that there exists a sequence $(u_n)_{n \in \mathbb{N}}$ with $\|u_n - u^*\|_U < \frac{\varepsilon}{2n}$ and $\mathcal{T}(u_n) \not\subseteq B(u^*, \varepsilon)$. Hence, there exists a sequence $(\tilde{v}_n)_{n \in \mathbb{N}}$ with $\tilde{v}_n \in \mathcal{T}(u_n)$ and $\|\tilde{v}_n - u^*\|_U \geq \varepsilon$. Because $\mathcal{T}(u_n)$ is path connected,

there exist continuous functions $c_n : [0, 1] \rightarrow \mathcal{T}(u_n)$ with $c_n(0) = u_n$ and $c_n(1) = \tilde{v}_n$. Because $\|c_n(0) - u^*\|_U < \frac{\varepsilon}{2}$ and $\|c_n(1) - u^*\|_U \geq \varepsilon$, the intermediate value theorem yields the existence of $v_n = c_n(\tau_n) \in \mathcal{T}(u_n)$ for some $\tau_n \in [0, 1]$ satisfying

$$(5) \quad \|v_n - u^*\|_U = \frac{\varepsilon}{2}.$$

By Theorem 12, we obtain for the distance to the limit v_n^* of the r -regular part of the generalized Newton path emanating from v_n that

$$\begin{aligned} \|v_n - v_n^*\|_U &\leq \ell(v_n^n) < \frac{r}{1-\kappa} \|F(v_n)\|_V \leq \frac{r}{1-\kappa} \|F(u_n)\|_V < \frac{r^2}{1-\kappa} \|f(u_n)\|_U \\ &= \frac{r^2}{1-\kappa} \|f(u_n) - f(u^*)\|_U \leq \frac{r^2 L}{1-\kappa} \|u_n - u^*\|_U < \frac{r^2 L \varepsilon}{2(1-\kappa)n} \rightarrow 0, \end{aligned}$$

which implies for some sufficiently large n that $v_n^* \in B(u^*, \varepsilon) \subseteq \mathcal{R}_r$ and thus $F(v_n^*) = 0$. By (5) we get $v_n^* \neq u^*$ in contradiction to the uniqueness of u^* . \square

We use the next lemma to prove discrete descent of the residual norm.

LEMMA 14. *A2 holds if and only if for all $u \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ and $t \in [0, 1]$*

$$\|F(u) - tF'(u)f(u)\|_V \leq [1 - (1-\kappa)t] \|F(u)\|_V.$$

Proof. As in [24], the nontrivial direction of the proof follows from the convexity of the functional $\varphi(t) = \|F(u) - tF'(u)f(u)\|_V$ and A2 according to

$$\varphi(t) \leq (1-t)\varphi(0) + t\varphi(1) \leq [(1-t) + \kappa t] \|F(u)\|_V. \quad \square$$

LEMMA 15. *Let A1, A2 and A3 hold. If $u_k \in \mathcal{R}_r \cap \mathcal{T}(u_0)$, then*

$$\|F(u_{k+1})\|_V \leq \left[1 - (1-\kappa)t_k + \frac{\omega}{2} \|f(u_k)\|_U t_k^2\right] \|F(u_k)\|_V.$$

Furthermore, if there exists a $\theta < 1$ such that the step size sequence satisfies

$$\omega t_k \|f(u_k)\|_U \leq 2\theta(1-\kappa),$$

then

$$\|F(u_{k+1})\|_V \leq [1 - (1-\theta)(1-\kappa)t_k] \|F(u_k)\|_V.$$

Proof. Using the fundamental theorem of calculus for functions with values in Banach spaces, Lemma 14, and A3 we obtain the first assertion from

$$\begin{aligned} \|F(u_{k+1})\| &= \left\| F(u_k) - \int_0^{t_k} F'(u_k - \tau f(u_k)) f(u_k) d\tau \right\|_V \\ &= \left\| F(u_k) - t_k F'(u_k) f(u_k) + \int_0^{t_k} [F'(u_k) - F'(u_k - \tau f(u_k))] f(u_k) d\tau \right\|_V \\ &\leq \|F(u_k) - t_k F'(u_k) f(u_k)\|_V + \int_0^{t_k} \| [F'(u_k) - F'(u_k - \tau f(u_k))] f(u_k) \|_V d\tau \\ &\leq \left[1 - (1-\kappa)t_k + \frac{\omega}{2} \|f(u_k)\|_U t_k^2\right] \|F(u_k)\|_V. \end{aligned}$$

The second assertion follows immediately. \square

We can now state a local convergence theorem.

THEOREM 16. Let A1, A2, and A3 hold. If there exists a $\bar{t} \in (0, 1)$ such that $t_k \geq \bar{t}$ for all $k \in \mathbb{N}$ and if there exists a $\theta \in (0, 1)$ such that for some $k \in \mathbb{N}$ the iterate u_k satisfies

$$\mathcal{T}(u_k) \subseteq \mathcal{R}_r \quad \text{and} \quad \omega r \|F(u_k)\|_V \leq 2\theta(1 - \kappa),$$

then $(u_k)_{k \in \mathbb{N}}$ converges to some point $u^* \in \mathcal{T}(u_k)$ with $F(u^*) = 0$.

Proof. Because $t_k \in [0, 1]$, we have

$$\omega t_k \|f(u_k)\|_U \leq \omega \|f(u_k)\|_U \stackrel{A1}{\leq} \omega r \|F(u_k)\|_V \leq 2\theta(1 - \kappa).$$

Hence, repeated application of Lemma 15 yields for all $j \in \mathbb{N}$

$$(6) \quad \|F(u_{k+j})\|_V \leq q^j \|F(u_k)\|_V \quad \text{with } q := 1 - (1 - \theta)(1 - \kappa)\bar{t}.$$

Because $q < 1$, $\|F(u_k)\|_V$ converges geometrically. In addition, we obtain that $(u_k)_{k \in \mathbb{N}}$ is a Cauchy sequence by virtue of

$$\begin{aligned} \|u_k - u_{k+j}\|_U &\leq \sum_{i=0}^{j-1} \|u_{k+i} - u_{k+i-1}\|_U = \sum_{i=0}^{j-1} t_{k+i} \|f(u_{k+i})\|_U \\ &\leq r \sum_{i=0}^{j-1} \|F(u_{k+i})\|_V \leq r \|F(u_k)\|_V \sum_{i=0}^{\infty} q^i = \frac{r}{1-q} \|F(u_k)\|_V \xrightarrow{k \rightarrow \infty} 0. \end{aligned}$$

Thus, $(u_k)_{k \in \mathbb{N}}$ converges to some $u^* \in \mathcal{T}(u_k) \subseteq \mathcal{R}_r$ and (6) implies $F(u^*) = 0$. \square

For the rate of convergence, we obtain the following result:

LEMMA 17. Under the assumptions of Theorem 16, $\|F(u_k)\|_V$ converges linearly with asymptotic linear convergence rate $\kappa < 1$.

Proof. *Proof.* Because $u_k \in \mathcal{R}_r$, it follows that $\|f(u_k)\|_U \leq r \|F(u_k)\|_V \rightarrow 0$. Hence, there is a sequence $(\theta_K)_{K \in \mathbb{N}}$ with $\theta_K \in [0, 1]$ and $\theta_K \rightarrow 0$ such that

$$\omega t_k \|f(u_k)\|_U \leq 2\theta_K(1 - \kappa) \quad \text{for all } k \geq K.$$

Lemma 6 and repeated application of Lemma 15 then deliver

$$\|F(u_{K+1})\|_V \leq [1 - (1 - \theta_K)(1 - \kappa)] \|F(u_K)\|_V,$$

where $1 - (1 - \theta_K)(1 - \kappa) \rightarrow \kappa$ as $K \rightarrow \infty$. \square

In order to obtain methods with guaranteed superlinear or quadratic local convergence, it is necessary to drive κ to zero as $F(u_k) \rightarrow 0$.

2.7. Global convergence. In order to prove an a priori bound on the decrease of the nonlinear residual for (BSC), we need the following Lemma.

LEMMA 18. Let $h > 0$. If a sequence $(a_k)_{k \in \mathbb{N}}$ of nonnegative numbers satisfies $a_{k+1}^2 \leq a_k^2 - 2ha_k$ for all $k \in \mathbb{N}$, then $a_k \leq \max\{a_0 - kh, h\}$ for all $k \in \mathbb{N}$.

Proof. The assertion is trivial for all $a_k \leq h$. Thus, let $k \geq 1$ with $a_{k-1} \geq h$. The assumption implies for all $j = 1, \dots, k$

$$a_j^2 \leq a_{j-1}^2 - 2ha_{j-1} \leq a_{j-1}^2 - 2ha_{j-1} + h^2 = (a_{j-1} - h)^2.$$

Because $a_{j-1} \geq a_{k-1} \geq h$, the last inequality implies $a_j \leq a_{j-1} - h$ and thus

$$a_j + jh - t \leq a_{j-1} + (j-1)h - t \quad \text{for all } t \in \mathbb{R}.$$

We now consider the nonnegative continuous function $\beta : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$\beta(t) := (\max\{t, 0\})^2 = \begin{cases} t^2 & \text{for } t \geq 0, \\ 0 & \text{for } t < 0. \end{cases}$$

Because $\beta(t)$ is nondecreasing, we obtain

$$\beta(a_j + jh - t) \leq \beta(a_{j-1} + (j-1)h - t) \quad \text{for all } t \in \mathbb{R},$$

which implies for $t = kh$ that

$$a_k^2 = \beta(a_k) \leq \beta(a_{k-1} - h) \leq \cdots \leq \beta(a_0 - kh) = (\max\{a_0 - kh, 0\})^2.$$

Taking the square root on both sides completes the proof. \square

THEOREM 19. *Let A1, A2, A3, A4, and A5 hold. Then there exists an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ iteration (2) with (BSC) satisfies exactly one of the following alternatives:*

1. *There exists a $k \in \mathbb{N}$ with $u_j \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ for all $j < k$ and $u_k \notin \mathcal{R}_r$.*
2. *For all $k \in \mathbb{N}$ the iterates satisfy $u_k \in \mathcal{R}_r \cap \mathcal{T}(u_0)$ and there exists an H -independent constant $c > 0$ for the a priori estimate*

$$\|F(u_k)\|_V^{\frac{1}{2}} \leq \max \left\{ \|F(u_0)\|_V^{\frac{1}{2}} - kc\sqrt{H}, c\sqrt{H} \right\} \quad \text{for all } k \in \mathbb{N}.$$

Proof. Let $\theta \in (0, 1)$. If $\omega = 0$, we choose $\eta > 0$ arbitrarily large. Otherwise, we set

$$\eta := \frac{2\theta(1-\kappa)}{\omega}.$$

Lemma 9 then yields the existence of an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ it holds that

$$t_k \|f(u_k)\| \leq \eta \quad \text{with } t_k = \min \mathcal{B}_H(u_k).$$

Hence, we can use Lemma 15 inductively for $k \in \mathbb{N}$ either until the first iterate $u_k \notin \mathcal{R}_r$ (alternative 1) or for $k \rightarrow \infty$ (alternative 2). The lower step size bound of Lemma 7 and Lemma 15 then yield

$$\begin{aligned} \|F(u_{k+1})\|_V &\leq [1 - (1-\theta)(1-\kappa)t_k] \|F(u_k)\|_V \leq \|F(u_k)\|_V - 2c\sqrt{H} \|F(u_k)\|_V^{\frac{1}{2}}, \\ \text{with } c &= \frac{(1-\theta)(1-\kappa)}{2\sqrt{rL}} > 0. \end{aligned}$$

Finally, Lemma 18 yields the a priori estimate with $a_k = \|F(u_k)\|_V^{\frac{1}{2}}$ and $h = c\sqrt{H}$. \square

The following Lemma is required for the final theorem, which assures convergence of the iterates to u_0^* .

LEMMA 20. *Let A4 hold. If $u^k(t_k + \tau), u^{k+1}(\tau) \in \mathcal{T}(u_k)$ for all $\tau \in [0, t]$, then*

$$\|u^k(t_k + t) - u^{k+1}(t)\|_U \leq \frac{1}{2} \|f(u_k)\|_U L e^{L(t_k+t)} t_k^2.$$

Proof. We use the integral form of the Gronwall inequality [24, Lemma 8.5]. \square

THEOREM 21. *Let A1, A2, A3, A4, and A5 hold. If $u_0^* \in \mathcal{R}_r$ is an isolated zero of F in a neighborhood of u_0^* , then there exists an $\overline{H} > 0$ such that for all $H \in (0, \overline{H}]$ iteration (2) with step size selection (BSC) converges to u_0^* .*

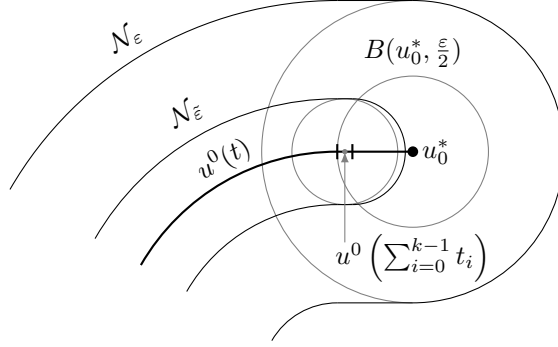


FIGURE 1. The idea of the proof of Theorem 21 is based on steering an iterate u_k into the $\tilde{\varepsilon}$ -ball around $u^0\left(\sum_{i=0}^{k-1} t_i\right)$, which is fully contained in the ε -ball around the solution u_0^* , the region of local convergence. The short black vertical dashes on the generalized Newton path $u^0(t)$ mark the points $u^0(T_* - 1)$ and $u^0(T_*)$.

Proof. From $u_0^* \in \mathcal{R}_r$ we obtain $u^0(t) \in \mathcal{R}_r$ for all $t \in [0, \infty)$ by Definition 11. Because \mathcal{R}_r is open, there exists an $\varepsilon > 0$ such that

$$\mathcal{N}_\varepsilon := \bigcup_{t \in [0, \infty)} B(u^0(t), \varepsilon) \subseteq \mathcal{R}_r.$$

Because u_0^* is an isolated zero of F , we can assume without loss of generality that $\varepsilon > 0$ was chosen small enough such that u_0^* is the only zero of F in \mathcal{N}_ε . Let now $\theta \in (0, 1)$. In anticipation of the final argument in this proof, we further assume without loss of generality ε to be sufficiently small to satisfy

$$(7) \quad \omega r^2 L \varepsilon \leq 2\theta(1 - \kappa) \quad \text{and, by Lemma 13,} \quad \mathcal{T}(u) \subseteq \mathcal{R}_r \text{ for all } u \in B(u_0^*, \varepsilon).$$

We now choose $T_* < \infty$ such that

$$\|u^0(t) - u_0^*\|_U \leq \frac{\varepsilon}{2} \quad \text{for all } t \geq T_* - 1.$$

Because u_0^* is the only zero of F on \mathcal{N}_ε , we can choose an $\tilde{\varepsilon} \in (0, \frac{\varepsilon}{2})$ such that there is a $\eta > 0$ satisfying

$$\|f(u)\|_U > \eta \quad \text{for all } u \in \mathcal{T}(u_0) \cap \mathcal{N}_{\tilde{\varepsilon}}, \text{ where } \mathcal{N}_{\tilde{\varepsilon}} := \bigcup_{t \in [0, T_*]} B(u^0(t), \tilde{\varepsilon}).$$

From A5 we obtain the existence of constants $\gamma, t_\gamma > 0$ such that

$$(8) \quad \|g(u, t)\|_U \geq \gamma t \quad \text{for all } t \in [0, t_\gamma] \text{ and } u \in \mathcal{N}_{\tilde{\varepsilon}}.$$

We can then use $\bar{t} = t_\gamma$ in Lemma 8 to obtain a constant $\bar{H} > 0$ such that

$$(9) \quad \min \mathcal{B}_H(u) \leq t_\gamma \quad \text{for all } H \in (0, \bar{H}] \text{ and } u \in \mathcal{N}_{\tilde{\varepsilon}}.$$

In anticipation of a later argument in the proof, we can assume without loss of generality that \bar{H} is sufficiently small to satisfy

$$(10) \quad T_* e^{LT_*} (rL \|F(u_0)\|_V)^{\frac{3}{2}} \bar{H}^{\frac{1}{2}} \leq 2\gamma\tilde{\varepsilon}.$$

Combining the inequalities (8) and (9) with (BSC), we see that

$$(11) \quad \gamma t_k^2 \leq t_k \|g(u_k, t_k)\|_U \leq H \quad \text{for all } H \in (0, \bar{H}] \text{ and } u_k \in \mathcal{N}_{\tilde{\varepsilon}}.$$

We now choose an H -dependent $\bar{k} \in \mathbb{N}$ that satisfies

$$T_* - 1 \leq \sum_{i=0}^{\bar{k}-1} t_i \leq T_*$$

and show by induction that $u_k \in \mathcal{T}(u_0) \cap \mathcal{N}_{\tilde{\varepsilon}}$ for all $k \leq \bar{k}$, which clearly holds true for $k = 0$ because $u_0 = u^0(0) \in B(u_0, \tilde{\varepsilon})$. We can now assume inductively that $u_i \in \mathcal{T}(u_0) \cap \mathcal{N}_{\tilde{\varepsilon}}$ for all $i \leq k-1$ with $k \leq \bar{k}$ in order to show $u_k \in \mathcal{T}(u_0) \cap \mathcal{N}_{\tilde{\varepsilon}}$. Because $\mathcal{N}_{\tilde{\varepsilon}} \subseteq \mathcal{R}_r$, Lemma 7 yields

$$T_* \geq \sum_{i=0}^{k-1} t_i \geq k \frac{\sqrt{H}}{\sqrt{rL \|F(u_0)\|_V}},$$

which implies the bound

$$(12) \quad k \leq \frac{T_* \sqrt{rL \|F(u_0)\|_V}}{\sqrt{H}}.$$

We then obtain by a telescope argument, Lemma 20, A1, (11), (12), and (10) that

$$\begin{aligned} \left\| u^0 \left(\sum_{i=0}^{k-1} t_i \right) - u_k \right\|_U &\leq \sum_{j=0}^{k-1} \left\| u^j \left(\sum_{i=j}^{k-1} t_i \right) - u^{j+1} \left(\sum_{i=j+1}^{k-1} t_i \right) \right\|_U \\ &\leq \sum_{j=0}^{k-1} \frac{1}{2} \|f(u_j)\|_U L e^{LT_*} t_j^2 < \frac{rL e^{LT_*} \|F(u_0)\|_V}{2\gamma} H k \\ &\leq \frac{T_* e^{LT_*} (rL \|F(u_0)\|_V)^{\frac{3}{2}}}{2\gamma} \sqrt{H} \leq \tilde{\varepsilon}. \end{aligned}$$

Because $\sum_{i=0}^{k-1} t_i \leq T_*$, we have established by induction that $u_k \in \mathcal{T}(u_0) \cap \mathcal{N}_{\tilde{\varepsilon}}$ for all $k \leq \bar{k}$. Finally, $u_{\bar{k}} \in B(u_0^*, \varepsilon)$ by virtue of

$$\|u_0^* - u_{\bar{k}}\|_U \leq \left\| u_0^* - u^0 \left(\sum_{i=0}^{\bar{k}-1} t_i \right) \right\|_U + \left\| u^0 \left(\sum_{i=0}^{\bar{k}-1} t_i \right) - u_{\bar{k}} \right\|_U < \frac{\varepsilon}{2} + \tilde{\varepsilon} < \varepsilon.$$

It follows from A1, A4, and (7) that

$$\begin{aligned} \omega r \|F(u_{\bar{k}})\|_V &< \omega r^2 \|f(u_{\bar{k}})\|_U = \omega r^2 \|f(u_{\bar{k}}) - f(u_0^*)\|_U \\ &\leq \omega r^2 L \|u_{\bar{k}} - u_0^*\|_U < \omega r^2 L \varepsilon \leq 2\theta(1 - \kappa). \end{aligned}$$

Thus, Theorem 16 establishes convergence to the unique zero u_0^* in $\mathcal{N}_{\tilde{\varepsilon}}$. \square

2.8. A note on generalizations to Banach spaces. The only step in the convergence proof that exploits the Hilbert space structure of V is Lemma 10. All remaining steps can be carried out even if V is only a Banach space. The general

approach here is to modify the used level function and to require an additional assumption akin to condition A2.

In this section, we shortly present the necessary modifications for the special case of the Lebesgue space $V = L^p(\Omega)$ of real-valued p -integrable functions with $2 < p < \infty$ and $\Omega \subset \mathbb{R}^n$. First, we need to consider a different level function

$$T_p(u) = \frac{1}{p} \|F(u)\|_V^p = \frac{1}{p} \int_{\Omega} |F(u)(s)|^p ds,$$

with analogously defined level sets $\mathcal{T}_p(u)$. In addition to A2, we require pointwise that

$$[F(u) - F'(u)f(u)](s) \leq \kappa |F(u)(s)| \quad \text{for almost all } s \in \Omega \text{ and all } u \in \mathcal{R}_r \cap \mathcal{T}_p(u_0).$$

Standard arguments on the differentiability of p -norms (see, e.g., [21, Thm. 2.6]) then deliver with the abbreviation $u = u^k(t)$ that

$$\begin{aligned} \frac{d}{dt} T_p(u^k(t)) &= - \int_{\Omega} |F(u)(s)|^{p-2} [F(u)(s) \cdot (F'(u)f(u))(s)] ds \\ &= - \int_{\Omega} |F(u)(s)|^p ds + \int_{\Omega} |F(u)(s)|^{p-2} [F(u)(s) \cdot (F(u) - F'(u)f(u))(s)] ds \\ &\leq -pT_p(u) + \int_{\Omega} |F(u)(s)|^{p-1} (F(u) - F'(u)f(u))(s) ds \\ &\leq -p(1 - \kappa)T_p(u) \leq 0. \end{aligned}$$

After application of Gronwall's inequality, multiplication by p and taking the p -th root, we establish the result of Lemma 10

$$\|F(u^k(t))\|_V \leq e^{-(1-\kappa)t} \|F(u_k)\|_V.$$

2.9. Algorithmic realization. The algorithmic realization of (BSC) can be carried over verbatim from the finite dimensional setting laid out in [24, §10] with the use of $\|\cdot\|_U$ for all occurring norms. As in [24], we do not use monotone iterations [14] for numerical computations here but use the bisection procedure with exponentially smoothed step size prediction. For convenience, we sketch it again: In order to determine t_k from (BSC), we approximately compute a zero of the Lipschitz continuous scalar function $t \mapsto t \|g(u_k, t)\|_U - H$ by a bracketing procedure. Numerically, we are content with a t_k that satisfies

$$(13) \quad t_k \|g(u_k, t_k)\|_U \in [H^l, H^u] \quad \text{or} \quad t_k = 1 \text{ and } t \|g(u_k, t)\|_U < H \text{ for all } t \in [0, 1],$$

where $H^l < H$ and $H^u > H$ are close to H .

The step size prediction in [24] is vital for minimizing the number of bracketing steps. If done correctly, the predicted step size often already satisfies (13) in all but a few iterations and thus almost no extra computational effort in terms of residual and increment evaluations $f(u_k)$ is required for the globalization procedure in most iterations.

3. Design of Newton-type methods. The convergence analysis of backward step control lends itself immediately to the design of globally convergent Newton-type methods. The two required steps are:

1. Define M (or directly f) respecting the κ -condition A2.
2. Use (BSC) to determine the step size sequence (t_k) .

The second step is generic. For the first step, we give two important examples in the following two sections. Both methods find approximations δu_k of the Newton increment $\delta u_k^{\text{Newton}}$ determined by the (infinite dimensional) linear system

$$(14) \quad F'(u_k)\delta u_k^{\text{Newton}} = -F(u_k).$$

We emphasize that the operator $M(u_k)$ is implicitly determined by requiring $\delta u_k = -M(u_k)F(u_k)$. It is not required to actually compute $M(u_k)$, as long as we have δu_k . However, we need to make sure that f is Lipschitz continuous with respect to u .

3.1. Krylov–Newton methods. Krylov subspace methods like GMRES [26] for the iterative solution of linear systems, originally developed for large but finite dimensional sparse systems, can also be stated for infinite dimensional linear operators. As expected, the convergence theory is more complicated in the infinite dimensional case (see, e.g., [22, 15]). If the structure of $F'(u_k)$ admits the application of GMRES, the m -th iterate δu_k^m of GMRES applied to (14) satisfies the minimum residual property

$$(15) \quad \delta u_k^m = \arg \min_{\delta u \in \mathcal{K}^m(F'(u_k), F(u_k))} \|F(u_k) + F'(u_k)\delta u\|_V,$$

where the (at most m -dimensional) m -th Krylov subspace is defined as

$$\mathcal{K}^m(F'(u_k), F(u_k)) = \{q(F'(u_k))F(u_k) \mid q \text{ is a polynomial of degree less than } m\}.$$

In the light of (15), the κ -condition A2 is nothing but the classical relative termination condition of GMRES with given relative tolerance κ . In other words, GMRES minimizes κ over the Krylov subspace and thus yields (with Lemma 17) an asymptotic linear convergence rate of κ for the nonlinear Krylov–Newton method. Thus, the backward step convergence theory of §2 delivers suitable termination criteria for the inner linear iterations, which are hard to find (c.f. the discussion in [25]) or purely heuristic for other nonlinear methods. In particular, a rather loose relative stopping criterion of, say, $\kappa = \frac{1}{10}$ delivers asymptotically already one decimal digit of accuracy per nonlinear iteration.

With this approach, there is one theoretic gap we need to close: The increment $-f(u_k) = -M(u_k)F(u_k) = \delta u_k^m$ depends on the number of GMRES iterations m , which is determined adaptively. As a concatenation of a finite number of Lipschitz continuous operations, the m -th GMRES iterate depends Lipschitz continuously on u_k , but changes in m from one nonlinear k -iteration to another can lead to discontinuities in the operator M . However, a small modification of the above approach can make sure that $M(u)$ and thus $f(u)$ are Lipschitz continuous with respect to u , as required for A4: Instead of using the final GMRES iterate δu_k^m , we could use a linear combination $f(u_k) = (1 - \alpha_k)\delta u_k^{m-1} + \alpha_k\delta u_k^m$ of the two last GMRES iterates such that instead of the inequality A2 the equality

$$\nu(\alpha_k) := \|F(u_k) + F'(u_k) [(1 - \alpha_k)\delta u_k^{m-1} + \alpha_k\delta u_k^m]\|_V = \kappa \|F(u_k)\|_V$$

holds for some $\alpha_k \in [0, 1]$. This is always possible by virtue of the intermediate value theorem applied to the continuous function $\nu(\alpha)$, which satisfies $\nu(0) > \kappa \|F(u_k)\|$ and $\nu(1) \leq \kappa \|F(u_k)\|$ because GMRES has terminated in step m but not yet in step $m - 1$.

Based on our experience in practical computations, however, it is more efficient to always use $\alpha_k = 1$ and weaken the bisection procedure for the approximate solution

to (BSC) against discontinuities of g by relaxing the lower bound H^1 in (13) closer to zero. This might give rise to smaller than necessary step sizes t_k , which has not been observed to be problematic in practical computations, but usually delivers faster local residual contraction once $t_k = 1$.

We report numerical results of a GMRES–Newton method for the Carrier equation in §5.1.

3.2. Approximation by discretization. We can also compute an approximate solution δu_k by first discretizing (14) and then (approximately) solving the discretized system. The κ -condition A2 yields a computable criterion for checking if the approximation is accurate enough to ensure convergence in U . If not, we need to improve the discretization (and possibly the accuracy of the approximate solution to the resulting finite dimensional linear system).

In this conceptually simple approach, challenges can arise in the evaluation of the V -norms in A2. We address these issues for the case of an elliptic partial differential equation in §5.2. Moreover, the evaluation of the V -norm in A2 can provide a means to adaptively discretize (14) and in turn also the original nonlinear problem.

As in §3.1, the procedure for approximating a solution to (14) usually involves some discrete decisions, for instance the marking and refinement of certain discretization cells as k increases. Thus, the so constructed $f(u)$ is not Lipschitz continuous. A smoothed formulation akin to the interpolation construction in §3.1 exceeds the scope of this article and shall be investigated in future work.

4. Example: Nonlinear elliptic boundary value problems. In this section, we illustrate the general paradigms presented in §3.2 for the class of elliptic boundary value problems on a bounded domain $\Omega \subset \mathbb{R}^n$ with continuously differentiable boundary $\partial\Omega$ that can be cast as nonlinear root-finding problems (1) with the Sobolev spaces $U = H_0^1(\Omega)$ and $V = H^{-1}(\Omega)$.

Based on the Poincaré inequality (see, e.g., [12]), we can use the inner product

$$(u, v)_U = \int_{\Omega} \nabla u \cdot \nabla v$$

for the Hilbert space U . For the Hilbert space V , we can then compute norms via the Riesz representation theorem [28, §III.3]: For every $v \in V$, there exists a uniquely determined $r_v \in U$ such that

$$(16) \quad (u, r_v)_U = \int_{\Omega} v u \text{ for all } u \in U \quad \text{and} \quad \|v\|_V = \|r_v\|_U.$$

4.1. Preconditioned GMRES. We first illustrate the general GMRES approach presented in §3.1. As a result, we obtain an appropriate preconditioner for GMRES purely based on the topology of U and V but independent of the problem instance at hand. As our focus in the case of Krylov–Newton methods here lies on a concise algorithmic statement rather than ultimate computational speed, we use Chebfun [5, 11] as an algorithmic tool for the numerical results in §5.1, because it allows to compute with functions (represented as adaptively truncated Chebyshev expansions) instead of numerical numbers [27] and supports the automatic computation of Fréchet derivatives by the use of automatic differentiation in function space [8].

Because the linear operators in Chebfun are implemented in strong form, we also use the strong form of the inner product in U

$$(u, v)_U = - \int_{-1}^1 u \Delta v = - \int_{-1}^1 v \Delta u,$$

(which requires u or v to have square integrable second derivatives). It follows from (16) that

$$-\int_{-1}^1 u \Delta r_v = (u, r_v)_U = \int_{-1}^1 v u \quad \text{for all } u \in U,$$

and, thus, $r_v = -\Delta^{-1}v$ and

$$\|v\|_V^2 = \|r_v\|_U^2 = \|-\Delta^{-1}v\|_U^2.$$

In Chebfun, Δ^{-1} can be evaluated efficiently using ultraspherical spectral collocation [23]. Based on these prerequisites, we modified Chebfun's builtin GMRES to use the inner product and norm of U (instead of $L^2(\Omega)$) in combination with Δ^{-1} as a preconditioner, which yields the correct residual norm $\|\Delta^{-1}v\|_U = \|v\|_V$. The numerical results for the Carrier equation in §5.1 indicate that this preconditioner works satisfactorily in practice.

4.2. Approximation by Finite Elements. In contrast to §4.1, we now explicitly discretize the increment δu_k and the step determination equation by Finite Elements: To this end, let \mathcal{C} be a partition of Ω into cells $C \in \mathcal{C}$. We can then construct the finite dimensional Finite Element subspace

$$U_{\mathcal{C}}^p = \{u \in C^0(\Omega) \mid u \text{ is a polynomial of degree } p \text{ on each } C \in \mathcal{C}\} \subset U.$$

The increment is then determined by finding $\delta u_k \in U_{\mathcal{C}}^p$ such that

$$(17) \quad (F'(u_k)\delta u_k)\varphi = -F(u_k)\varphi \quad \text{for all } \varphi \in U_{\mathcal{C}}^p.$$

By fixing a basis of $U_{\mathcal{C}}^p$, we obtain a linear system with a large but sparse (typically symmetric positive definite) $|U_{\mathcal{C}}^p|$ -by- $|U_{\mathcal{C}}^p|$ matrix.

4.2.1. Computation of norms in V . The discretization in the previous paragraph is completely standard. We now describe the non-standard part, which comprises the computation of

$$(18) \quad \|F(u_k)\|_V \quad \text{and} \quad \kappa_k := \frac{\|F(u_k) + F'(u_k)\delta u_k\|_V}{\|F(u_k)\|_V}$$

in the space $V = H^{-1}(\Omega)$. To this end, we use again the Riesz representation (16). However, using the same Finite Element subspace $U_{\mathcal{C}}^p$ for the discretization of (16) would yield the wrong value $\kappa_k = 0$ because the numerator vanishes. Moreover, using $U_{\mathcal{C}}^p$ would also give wrong results for $\|F(u_k)\|_V$ because the residual projected on $U_{\mathcal{C}}^p$ converges locally quadratic (as a Newton method on a finite dimensional space) if we solve (17) exactly, but does not see the discretization error. Thus, we need to solve (16) on richer Finite Element spaces. The numerical results in §5.2 indicate that choosing $U_{\mathcal{C}}^{p+1}$ seems to be sufficient for good estimates of the required V -norms.

4.2.2. Adaptive mesh refinement to minimize the contraction rate κ . Using p -refinement for the solution of (16) instead of refinement of \mathcal{C} has two advantages: First, the increase in the degrees of freedom $|U_{\mathcal{C}}^{p+1}|$ with respect to $|U_{\mathcal{C}}^p|$ is only moderate if p is moderately large, e.g., $p = 3$. Second, the squared norm $\|r_v\|_U^2$ can then be written as a sum of contributions from each cell $C \in \mathcal{C}$, which indicate which cells should ideally be refined if κ_k is larger than a desired residual contraction rate $\kappa < 1$ prescribed by the user. The cellwise contributions κ_k can be treated in the same fashion as existing cellwise error indicators. Of course, refining the mesh cells would also be possible but would require more notoriously slow operations on the mesh.

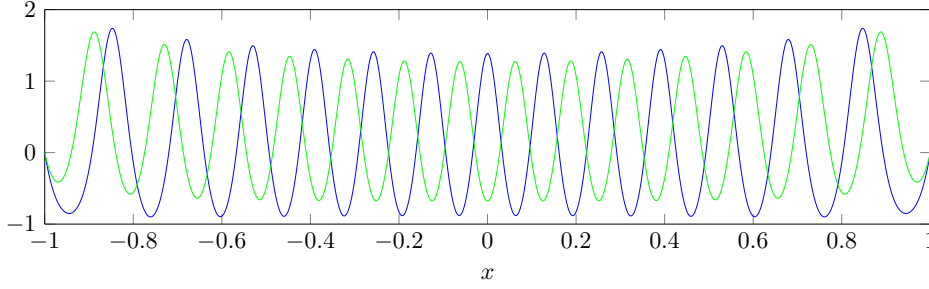


FIGURE 2. The local solutions to the Carrier equation with $\varepsilon = 10^{-3}$ obtained by a GMRES–Newton method with backward step control for $\kappa = 10^{-2}$ and varying values of H_{rel} (0.1: green, 0.05 and 0.01: blue).

5. Numerical results. We investigate the numerical performance of backward step control on two nonlinear elliptic boundary value problems on bounded domains Ω with continuously differentiable boundary. For the correct topology of the function space setting, we choose the Sobolev spaces $U = H_0^1(\Omega)$ and $V = H^{-1}(\Omega)$ as in §4.

All algorithmic parameters of backward step control are chosen as in [24] unless otherwise stated.

5.1. Carrier equation. For $\varepsilon > 0$, we want to determine a real-valued function $u(x)$ on $x \in [-1, 1]$ that solves the nonlinear second order boundary value problem

$$(19) \quad \varepsilon \Delta u + 2(1 - x^2)u + u^2 = 1, \quad u(\pm 1) = 0,$$

which—according to [7, §9.7]—is due to Carrier. For small ε , it becomes challenging to solve (19) because of boundary layers and the existence of many local solutions (compare Fig. 2).

We apply the GMRES–Newton method described in §4.1 to (19) and illustrate the convergence of the method in Fig. 3 for $\varepsilon = 10^{-3}$, $\kappa = 10^{-2}$, and varying values of $H = H_{\text{rel}} \|\delta u_0\|_U$, where we choose $H_{\text{rel}} \in \{0.5, 0.1, 0.05, 0.01\}$. The initial guess is $u_0 = 0$ and the termination criterion is $\|F(x_k)\|_V \leq 10^{-11}$. The relative GMRES termination tolerance κ was chosen rather large but at the same time small enough to ensure sufficiently fast local convergence with two decimal digits per iteration. We first observe that no convergence can be obtained for $H_{\text{rel}} = 0.5$, even though $\|F(u_{41})\|_V$ drops below $2.6 \cdot 10^{-5}$ and $\|\delta u_7\|_U \approx 0.28$. From these numbers we can estimate the nonlinearity of the problem in terms of ω and its well-posedness in terms of r based on Lemma 15, which yields

$$\omega \geq 2(1 - \kappa)/(t_7 \|f(u_7)\|_U) \approx 7.1 \quad \text{and} \quad \omega r \geq 2(1 - \kappa)/(t_{41} \|F(u_{41})\|_V) \approx 7.6 \cdot 10^4$$

and shows that the problem is highly nonlinear.

For the remaining choices of H_{rel} we obtain convergence, albeit a different local solution is found for $H_{\text{rel}} = 0.1$ than for the others (compare Fig. 2), which nicely confirms the theory of Theorem 21. As guaranteed by Lemma 6, full steps $t_k = 1$ are taken in the vicinity of a solution and we can clearly observe the asymptotic linear convergence rate of $\kappa = 10^{-2}$ for the residual norm predicted by Lemma 17. The final increment norms seem rather large because we do not compute δu_k if $\|F(u_k)\|_V$ is already below 10^{-11} . Thus, the last increment norm lags behind by one iteration and would be much smaller if we computed it again for the final iterate.

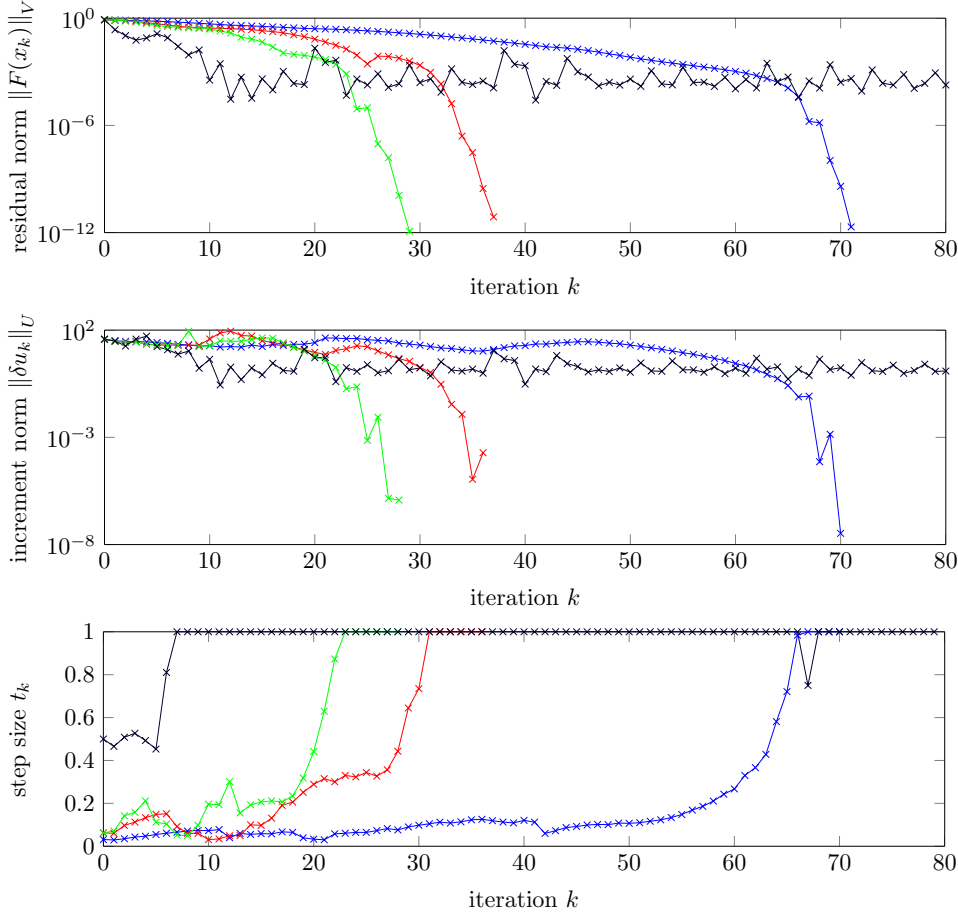


FIGURE 3. The residual and increment norms and the step size sequence for a GMRES–Newton method with backward step control applied to the Carrier equation with $\varepsilon = 10^{-3}$ for $\kappa = 10^{-2}$ and varying values of H_{rel} (0.5: black, 0.1: green, 0.05: red, 0.01: blue).

In Fig. 4, we see that the number of GMRES iterations needed in each nonlinear iteration stays modestly small. In each GMRES iteration, one operator-vector-multiplication must be carried out, which we compute via Chebfun as a directional derivative of F . In total, 1255 ($H_{\text{rel}} = 0.1$), 1455 ($H_{\text{rel}} = 0.05$), and 2471 ($H_{\text{rel}} = 0.01$) directional derivatives of F are required to solve the problem.

In addition, we can observe from Fig. 4 that the predicted step size t_k needs to be corrected only in few iterations k by the backward step control bisection procedure outlined in §2.9. In the first iteration, four ($H_{\text{rel}} = 0.1, 0.05$) and five ($H_{\text{rel}} = 0.01$) bisection steps are required to reduce the initial step size guess $t_0 = 1$. In all other iterations, only one additional bisection step is necessary (marked by \circ in Fig. 4), except in iteration $k = 8$ for $H_{\text{rel}} = 0.1$ and $k = 21$ for $H_{\text{rel}} = 0.01$, where two bisection steps need to be taken. This backs up our claim that the computational overhead of backward step control is small.

5.2. Minimum surface equation. In this section, we consider the classical minimum surface problem in the following special form: Let $\Omega = \{x \in \mathbb{R}^2 \mid \|x\|_2 < 1\}$

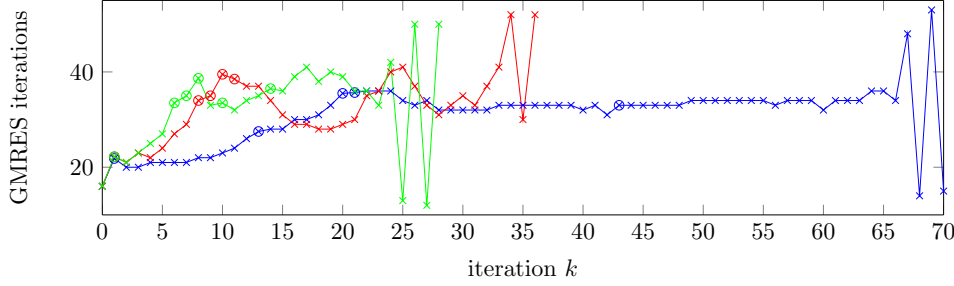


FIGURE 4. The average number of GMRES iterations for a GMRES-Newton method with backward step control for $\kappa = 10^{-2}$ and varying values of H_{rel} (0.1: green, 0.05: red, 0.01: blue) for the Carrier equation with $\varepsilon = 10^{-3}$. In the iterations marked with \circ , the backward step control bisection procedure performs extra iterations to determine t_{k-1} and δu_k and the plotted GMRES iterations are averaged over the number of bisection steps.

and $u^\partial(x) = \sin(2\pi(x_1 + x_2))$. We seek a function u on Ω that equals u^∂ on $\partial\Omega$ and minimizes the area of its graph

$$\min I(u) = \int_{\Omega} \sqrt{1 + |\nabla u|^2} \quad \text{s.t.} \quad u|_{\partial\Omega} = u^\partial|_{\partial\Omega}.$$

With the spaces $U = H_0^1(\Omega)$ and $V = H^{-1}(\Omega)$ as before, the minimum is described as the solution $u \in u^\partial + U$ to the variational problem

$$F(u)\varphi := \int_{\Omega} a(u) \nabla \varphi \cdot \nabla u = 0 \quad \text{for all } \varphi \in U, \quad \text{where } a(u) := \left(1 + |\nabla u|^2\right)^{-\frac{1}{2}}.$$

Thus, F maps $u^\partial + U$ to V . Its Fréchet derivative $F' : u^\partial + U \rightarrow \mathcal{L}(U, V)$ can then be expressed as

$$(F'(u)\delta u)\varphi = \int_{\Omega} a(u) \nabla \varphi \cdot (1 - a^2(u) \nabla u \otimes \nabla u) \nabla \delta u.$$

We solve the resulting system (17) only approximately with a preconditioned Conjugate Gradient (PCG) method [17]. The resulting inexactness also contributes to the computations of κ_k in (18).

All computations were obtained with the software package deal.II [3, 4].

5.2.1. Efficient computation of Riesz representations. As described in §4.2, the Riesz representations r_v need to be computed from (16) in order to compute the V -norms entering κ_k . The following algorithmical and computational approaches are important to prevent the computation times for the solution of (16) on the richer Finite Element subspace $U_{\mathcal{C}}^{p+1}$ from dominating the overall computational effort:

1. We found that PCG with a symmetric Gauss-Seidel smoother as preconditioner delivers good results. In the computations reported below, we employed a Chebyshev smoother of degree four because it yields a slightly better performance than Gauss-Seidel when running the code on several processors in parallel. The use of multigrid does not pay off because it usually involves a rather expensive setup machinery on unstructured meshes (see, e.g., [18]) and the right-hand sides of (16) consist mainly of high-frequency residuals in $U_{\mathcal{C}}^{p+1} \setminus U_{\mathcal{C}}^p$, the low frequency residuals having been mostly eliminated on $U_{\mathcal{C}}^p$ already.

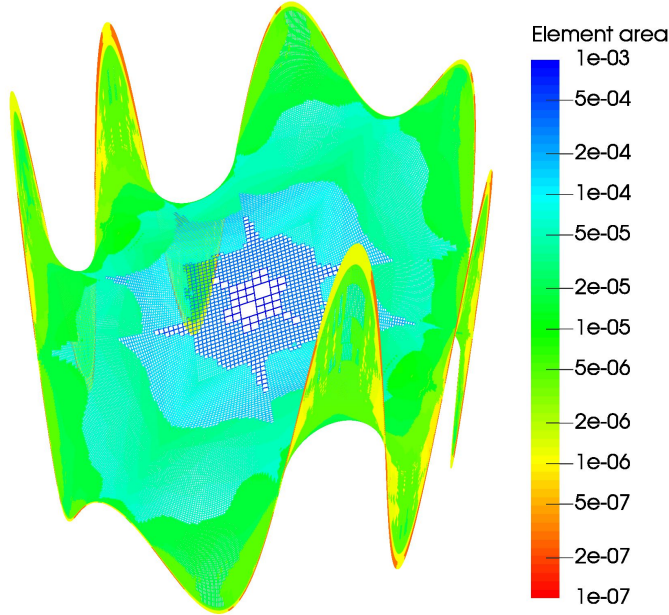


FIGURE 5. The final Finite Element mesh for the solution of the minimum surface equation with κ -optimizing adaptive mesh refinement. The element areas are encoded in color. Refinement is mostly necessary near the boundary, while large cells remain unrefined in the center.

2. In order to avoid a possible memory bottleneck caused by storing the stiffness matrix discretized on the high-dimensional space U_C^{p+1} , we use a matrix-free realization of the Laplacian [20].

3. Because the resulting computation times are then dominated by the bandwidth of the access to main memory, we perform all computations involved in the (matrix-free) matrix-vector products for the solution of (16) only with single instead of double precision floating point arithmetic. The numerical results below indicate that this approach is still sufficiently accurate, while being considerably faster.

4. Because κ_k only steers the algorithm but does not affect the quality of the iterates x_k directly, it can be computed with rather low accuracy requirements in the PCG method. We use relative stopping criteria of 0.1 and 0.05 for the numerator and the denominator of κ_k in (18).

5.2.2. Details about the numerical setup. We discretize (17) and (16) by nodal Finite Elements of order $p = 3$ and $p + 1 = 4$, respectively, on quadrilaterals with tensor product polynomials using Gauss-Lobatto nodes. For the elements on the curved boundary, we employ polynomial transformations of degree seven from the reference cell to the physical cells. The relative tolerance of PCG for the solution of (17) is 0.001 (in the Euclidean norm on the discretized vectors).

As a starting guess, we let u_0 be the Finite Element interpolation of u^∂ on the coarsest mesh depicted in Fig. 6. In a first phase, we iterate until the increment norm in U is below 0.01 without computing any V -norms. This first phase is a finite dimensional Newton method ($\kappa = 0$) globalized with backward step control ($H_{\text{rel}} = 0.05$, $t_0 = 1$) on the Finite Element subspace U_C^p belonging to the initial mesh.

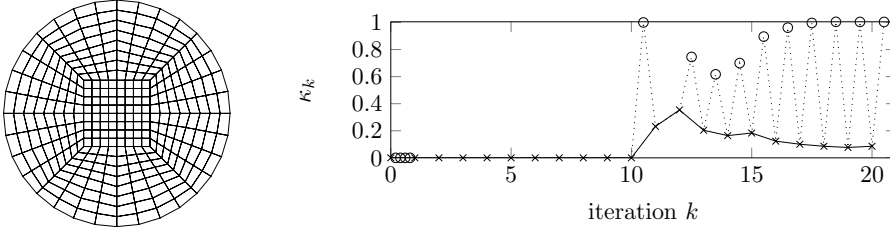


FIGURE 6. Left: The initial mesh for the solution of the minimum surface equation has 425 cells. With $p = 3$, the resulting Finite Element space has 2929 degrees of freedom. Right: The accepted and discarded values of κ_k in the numerical solution of the minimum surface equation. Discarded values are marked with \circ at fractional iteration numbers, the values of κ_k on the finally successful mesh are marked with \times at integer iteration numbers. The discarded values of κ_k converge to 1, while the accepted ones approach approximately 0.08.

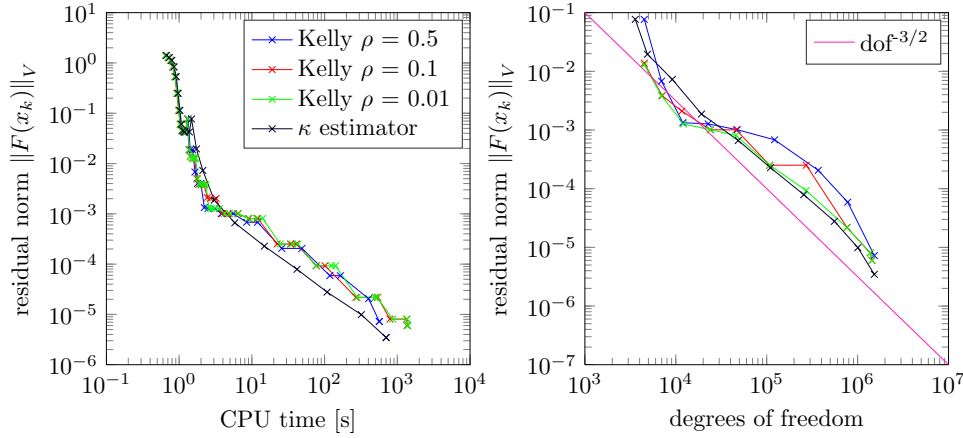


FIGURE 7. Convergence of the residuals for backward step control κ -optimizing mesh refinement (black) and mesh refinement based on the Kelly indicator for varying values of the required reductions $\rho = 0.5$ (blue), $\rho = 0.1$ (red), and $\rho = 0.01$ (green) on the current mesh for the minimum surface equation.

For the successive phase of nonlinear adaptive mesh refinement (see §4.2), we choose $\kappa = 0.5$. If $\kappa_k > \kappa$, we mark all cells for refinement that have a contribution of more than 2^{-p} times the maximum cell contribution to κ_k , up to a given maximum number of 200,000 cells. The resulting number of cells might be slightly higher in the final mesh due to mesh smoothing in deal.II. The numerical solution of the minimum surface equation is depicted in Fig. 5.

In the first eight iterations, the step size is gradually increased from $t_0 = 0.0625$ to $t_7 = 0.9738$. All other iterations are performed with full steps $t_k = 1$. The mesh is refined for the first time in iteration 11, kept for iteration 12, and then successively refined in each further step until the maximum number of cells is reached in iteration 21. We can furthermore observe from Fig. 6 that from iteration 16 on, the first trial value of κ_k before the refinement is always one (up to three decimal digits), which shows that no further improvement can be achieved by performing more nonlinear iterations on the current discretization, which is automatically detected correctly by the algorithm.

We compare in Fig. 7 the convergence of the residual norm $\|F(u_k)\|_V$ (computed

afterwards with high accuracy on a refined mesh, which is generated by one additional global refinement step of the triangulation of the final mesh) of backward step control κ -optimizing adaptive mesh refinement with the convergence when using mesh refinement with the deal.II builtin Kelly error indicator [19, 13]. In contrast to the theory of backward step control, there is no theoretical guideline for the Kelly indicator on how many nonlinear iterations to run before another round of refinement is triggered. We choose to refine as soon as the increment norm $\|\delta u_k\|_U$ becomes less than or equal to a factor $\rho > 0$ of the error estimate returned on the last mesh by the Kelly indicator. Fig. 7 shows that κ -optimizing adaptive mesh refinement delivers the best ratio of residual norm versus CPU time and versus the number of degrees of freedom compared to mesh refinement based on the Kelly indicator for varying values of $\rho = 0.5, 0.1, 0.01$.

The computations for the solution of the minimum surface equation with a final number of 1.5 million degrees of freedom using κ -optimizing mesh refinement took 112s wall clock time on the four cores of a mid 2012 MacBook Pro, 2.3 GHz Intel Core i7, 8 GB. Out of this grand total, the computations necessary for estimating κ_k took only 18s, even though they need to be performed on the high-dimensional Finite Element space U_C^{p+1} .

6. Conclusions. We presented a comprehensive convergence analysis for (BSC), a method that globalizes the convergence of Newton-type methods (2) for the solution of (1) in a Hilbert space setting. We proved that under the reasonable assumptions A1–A5 the iterates u_k either leave the region of r -regular points \mathcal{R}_r (in which case we need to adjust M in order to prevent attraction to singularities) or converge to the distinctive solution u_0^* (the initial guess u_0 propagated by the generalized Newton flow (3)) provided that $H > 0$ is chosen sufficiently small. Moreover, we provided an H -dependent a priori bound on the decrease of $\|F(u_k)\|_V$ and characterized the asymptotic linear residual convergence rate by κ . We provided efficient numerical methods based on the blueprint of bounding and optimizing κ in each iteration, either over a Krylov subspace in a GMRES–Newton method or through an adaptive Finite Element discretization, in order to balance the linearization error and the inexactness in the solutions of the linear systems. We applied these methods to the class of nonlinear elliptic boundary value problems and presented numerical results for the Carrier equation in a Chebfun implementation and for the minimum surface equation in deal.II, for which κ -optimizing adaptive mesh refinement delivers better performance than adaptive mesh refinement based on the Kelly estimator while still being totally generic in the sense that no problem dependent dual problems need to be set up. The only challenge here is to compute norms in $V = H^{-1}(\Omega)$ efficiently, which can be addressed by suitable numerical methods and techniques.

REFERENCES

- [1] M. Ainsworth and J.T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.
- [2] H. Amann. *Ordinary differential equations*, volume 13 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1990.
- [3] W. Bangerth, D. Davydov, T. Heister, L. Heltai, G. Kanschat, M. Kronbichler, M. Maier, B. Turcksin, and D. Wells. The deal.II library, version 8.4. *Journal of Numerical Mathematics*, 24, 2016.
- [4] W. Bangerth, R. Hartmann, and G. Kanschat. deal.II – a general purpose object oriented finite element library. *ACM Trans. Math. Softw.*, 33(4):24/1–24/27, 2007.

- [5] Z. Battles and L.N. Trefethen. An extension of MATLAB to continuous functions and operators. *SIAM J. Sci. Comput.*, 25(5):1743–1770, 2004.
- [6] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.*, 10:1–102, 2001.
- [7] C.M. Bender and S.A. Orszag. *Advanced mathematical methods for scientists and engineers. I.* Springer-Verlag, New York, 1999. Asymptotic methods and perturbation theory, Reprint of the 1978 original.
- [8] A. Birkisson and T.A. Driscoll. Automatic Fréchet differentiation for the numerical solution of boundary-value problems. *ACM Trans. Math. Software*, 38(4), 2012.
- [9] H.G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, volume 183 of *Bonner Mathematische Schriften*. Universität Bonn, Bonn, 1987.
- [10] P. Deuffhard. *Newton methods for nonlinear problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2011. Affine invariance and adaptive algorithms.
- [11] T.A. Driscoll, N. Hale, and L.N. Trefethen, editors. *Chebfun guide*. Pafnuty Publications, Oxford, 2014.
- [12] L.C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.
- [13] J.P. de S.R. Gago, D.W. Kelly, O.C. Zienkiewicz, and I. Babuška. A posteriori error analysis and adaptive processes in the finite element method. II. Adaptive mesh refinement. *Internat. J. Numer. Methods Engrg.*, 19(11):1621–1656, 1983.
- [14] A. Galántai and J. Abaffy. Always convergent iteration methods for nonlinear equations of Lipschitz functions. *Numer. Algor.*, pages 1–11, 2014.
- [15] M.G. Gasparo, A. Papini, and A. Pasquali. Some properties of GMRES in Hilbert spaces. *Numer. Func. Anal. Opt.*, 29(11–12):1276–1285, 2008.
- [16] T. Grätsch and K.-J. Bathe. A posteriori error estimation techniques in practical finite element analysis. *Comput. & Structures*, 83(4-5):235–265, 2005.
- [17] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952.
- [18] B. Janssen and G. Kanschat. Adaptive multilevel methods with local smoothing for H^1 - and H^{curl} -conforming high order finite element methods. *SIAM J. Sci. Comput.*, 33(4):2095–2114, 2011.
- [19] D.W. Kelly, J.P. de S.R. Gago, O.C. Zienkiewicz, and I. Babuška. A posteriori error analysis and adaptive processes in the finite element method. I. Error analysis. *Internat. J. Numer. Methods Engrg.*, 19(11):1593–1619, 1983.
- [20] M. Kronbichler and K. Kormann. A generic interface for parallel cell-based finite element operator application. *Comput. & Fluids*, 63:135–147, 2012.
- [21] E.H. Lieb and M. Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2001.
- [22] O. Nevanlinna. *Convergence of iterations for linear equations*. Lectures in mathematics. Birkhäuser, Basel, Boston, Berlin, 1993.
- [23] S. Olver and A. Townsend. A fast and well-conditioned spectral method. *SIAM Rev.*, 55(3):462–489, 2013.
- [24] A. Potschka. Backward step control for global Newton-type methods. *SIAM J. Numer. Anal.*, 54(1):361–387, 2016.
- [25] R. Rannacher and J. Vihharev. Adaptive finite element analysis of nonlinear problems: balancing of discretization and iteration errors. *J. Numer. Math.*, 21(1):23–61, 2013.
- [26] Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7(3):856–869, 1986.
- [27] L.N. Trefethen. Computing numerically with functions instead of numbers. *Math. Comput. Sci.*, 1(1):9–19, 2007.
- [28] K. Yosida. *Functional analysis*. Springer-Verlag, fifth edition, 1978.